



Polytechnic University of Puerto Rico  
Graduate School  
Course Schedule Fall 2009



Graduate Program in COMPUTER SCIENCE

PROG	CODE/SECT.	PRE-REQ.	COURSE	DAY	TIME	FACULTY	ROOM
Core Course	CECS 6030/31	Discrete Structure, Calculus II	Computational Theory	Sat	8:00 – 12:00 m	Dr. J. Duffany	L 301
	CECS 6750/ 35	UG OOP	Software Testing	Sat	1:00 – 5:00 pm	Dr. J. Valles	L 301
Elective Courses	CECS 6240/22	None	Technology Based Start-Up	Tues	6:30 - 10:30 pm	Dr. J. Ramirez	L 210
	CECS 6605/24	None	Database Systems	Thu	6:30 – 10:30 pm	Dr. A. Cruz	L 301
	CECS 7010/23	Calculus II	Computer Graphics I	Wed	6:30 – 10:30 pm	Dr. E. Lozano	L 301
	CECS 6824A/35	None	Spec. Topics in ITMIA (Web Spam and Internet Vulnerabilities: AIRWeb)	Sat	1:00 – 5:00 pm	Dr. E. Garcia	L 301
	EE 6150/21	UG OOP	Object Oriented Design	Mon	6:30 – 10:30 pm	Dr. E. Sobrino	L 302
	CECS 6760/22	EE 6130 or Coordinator Approval	Internet Engineering I	Tue	6:30 – 10:30 pm	Dr. O. Rodriguez	L 302
	CECS 7520/24	UG OOP	Human Computer Interaction	Thu	6:30 – 10:30 pm	Dr. O. Rodriguez	L 302

# Adversarial Information Retrieval on the Web

## A Graduate Course on Web Spam and Internet Vulnerabilities

*“So, You Want to know about:*

*Search Engines Spam?*

*Click-Through Frauds?*

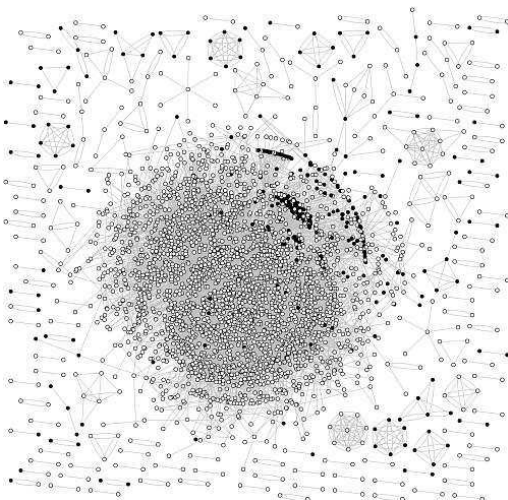
*SEO Snakeoil?*

*Email Spam & Exploits?*

*Malicious Web Crawlers?*

*Link Farms & Link Bombs?*

*If so, this Course is for You.”-Dr. E. Garcia.*



Since 2005, the AIRWeb Workshops have been part of either the SIGIR or W3C Conferences. During AIRWeb2007 a spam competition was celebrated, with a reference collection of Web pages, in which over 3,000 hosts were labeled by a team of volunteers as spam or non-spam. The picture at the left is a partial view of the corpus (black nodes are spam, white nodes are non-spam). Zoom in with your browser and try to identify all kind of link spam structures (reciprocal, triangular swapping, honey pots, etc).

Source:

<http://www.iw3c2.org/blog/2007/01/10/researching-web-spam/>

## **Title: Web Spam and Internet Vulnerabilities: AIRWeb (Adversarial Information Retrieval on the Web)**

**Description:** Commercial search engines like Google and Yahoo! are at the center of the Web as a connected graph, generating traffic to zillion of websites relevant to specific searches. This motivates content providers to try to do whatever it takes to rank highly in search engine result pages (SERPs). Such methods typically include dubious search engine optimization (SEO) and fraudulent search engine marketing (SEM) practices, malicious social networking, manipulation of link structures, and all kind of *spamdexing* techniques. Some of these techniques have been adopted by email spammers and computer hackers in an effort to find and exploit Internet Vulnerabilities. Collectively, these practices are known as Adversarial Information Retrieval. The material to be covered in this course is based on research papers presented at the AIRWeb Workshops. Students will be exposed to state-of-the-art and cutting-edge research. Students interested in conducting research on adversarial retrieval or whose research work is at the intersection of information security are encouraged to take this course.

**Target:** Students in Business, Engineering, and Computer Sciences and from other disciplines are encouraged to register for this special course.

**Requirements:** Permission from advisor or department.

**Grading:** Homeworks, Partial Exam, and a Final Exam.

**Topics:** Although not necessarily in this order, some of the topics to be covered, include, but are not limited to the followings:

- Web Crawlers and Email Crawlers
- Web-based Vulnerabilities
- E-Mail Spam
- Social Network Abuses and Exploits
- Server and Browser-based Exploits
- Link Bombing (a.k.a. Google-Bombing)
- Link Farms and Link Swapping Structures
- Click-Through Fraud and Spurious Web Analytics
- Comment Spam and Blog Spam
- Link & Spam Injections
- Malicious Tagging
- Reverse-Engineering of Ranking Algorithms
- Search Engine Optimization Spam
- Search Marketing Spam

**Textbook:** There is no official textbook. All lecture material is based on research work presented at the AIRWeb Workshops. This syllabus is subject to changes. Additional references and an extended syllabus will be provided in class. Syllabus, lecture plans, announcements, QA notes, etc will be provided online in the AIRWeb Course category of <http://irthoughts.wordpress.com> blog.

### **About Dr. Garcia**

Dr. Garcia research interests include Web Mining, Search Engine Architectures, and Information Retrieval at the intersection of Information Security and Intelligence. He is a program committee member of W3C's Adversarial Information Retrieval on the Web Workshops (AIRWeb), has served as reviewer for *JASIST*, *IBM's Computer and Graphics*, and has co-chaired several local conferences on search engine technologies. At Polytechnic University, he is a visiting lecturer, having taught the graduate courses *Web Mining & Business Intelligence* and *Search Engines Architecture*. He also conducts a research project on remote searches at Interamerican University of Puerto Rico, Metropolitan Campus. He is the founder of <http://www.miislita.com>, an online resource on information retrieval and search engine technologies.